An integrative platform to capture the orchestration of gesture and speech

Christelle Dodane, Dominique Boutet, Ivana Didirkova, Fabrice Hirsch, Slim Ouni, Aliyah Morgenstern

An integrative platform to capture the orchestration of gesture and speech

Abstract

A number of studies have highlighted the coordination of gesture and intonation (Bolinger, 1983; Darwin, 1872; Kendon, 1980) but the technological set-ups have been insufficient to couple the acoustic and gestural data with sufficient detail. In this paper, we present the MODALISA platform which enables language specialists to integrate gesture, intonation, speech production and content. The methods of data acquisition, annotation and analysis are detailed. The preliminary results of our pilot study illustrate strong correlations between gestures and intonation when they are simultaneously performed by the speaker. The correlations are particularly strong for proximal segments. Our aim is to expand those results and analyse typical and atypical populations across the lifespan.

1 Introduction

According to Bolinger (1983: 157), "*we READ intonation the same way we read gestures*". In parallel with Darwin's observations about gestures (1872), intonation is iconic in the sense that the meaning of upward and downward movements is related to attitudes and indirectly to metaphorical associations with tension, incompleteness and their opposites. Intonation has its own "*symbolizing power thanks to a primitive drive mechanism that raises pitch as tension rises and lowers it as tension falls*" (Bolinger, 1983: 156). It is part of our body movements which are more or less automatically concomitant to our state and our emotions. Bolinger highlights that gestures are coupled with intonation and display the same ascending and descending movements. Gesture and intonation may not systematically be produced together, but when they are, they are synchronized and co-expressive. Their synchrony does not necessarily mean that they work in unison, but rather that the parallel movements are coupled while the non-parallel movements are not. Several other authors have also noted a synchronization between the speech flow and the gestural flow. Kendon (1980: 211) states that "*it is as if the speech production process is manifested in two forms of activity simultaneously: in the vocal organs and also in bodily movement*". For example, the gestural stroke aligns temporarily with the specific linguistic segments that are co-expressive with it (McNeill, 1992). For these authors, gestures and speech are part of the same system. We also know that multimodal processes appear early on during language development, since canonical babbling is linked to the rhythmic and manual activities of babies (Locke *et al*., 1995). In adults, speech is often accompanied by gestures (Guellaï *et al*., 2014) and even congenital blind people gesture when interacting (Iverson et Goldin-Meadow, 1998). Adult speakers coordinate their gestural behaviors and intonation when they speak, both in terms of time and direction: downward / forward movements are typically produced with descending contours and upward / backward movements with ascending contours (Bolinger, 1983; Cruttenden, 1997). Balog and Brentari (2008) observed the same type of synchronization in children aged 12 to 24 months and showed that children coordinate their verbal and non-verbal behaviors at the temporal and directional levels as early as the first word period, in order to be better understood by those around them. In their study, gesture coding was done by hand by observers who used a video in slow-motion and they had to indicate whether there was synchronization with the intonation or not. Using motion capture (OptiTrak recordings) on ten speakers, Roustan and Dohen (2010) showed that the prosodic focus attracts the manual gestures (pointing, beat and control gestures), pointing gestures being the most synchronized gestures (mainly between the apex of the pointing gesture and articulatory vocalic targets). These studies indicate that it is crucial to work on the synchronization of prosody and gestural behaviors, in adults as well as children. In order to achieve that goal, the MODALISA team has planned to create a multimodal platform that will make it possible to analyze prosody and gesture together. Indeed, to our knowledge, there is no adequate instrument that makes it easy to measure gestures and prosody together. The objective of the

MODALISA[1] project is thus to create an integrative procedure with the existing tools, that would make it possible to align the acoustic data with the gestural data. Instead of manual coding, we aim to use automatic extractions of the different gestural components (movements of the hands, forearms and arms) using several motion capture systems. The original contribution of our project is that we use the gestural data complemented with articulatory and respiratory data obtained with other devices (laryngograph, articulograph, abdominal belt). This set-up allows us to create a truly multimodal platform for the simultaneous study of speech and gesture. It gives us access to objective, accurate and reliable data that will allow us to develop a large number of studies on speech and gesture. This paper presents our pilot study with the integrative system, our methodological procedure, preliminary results and perspectives.

2 Methods of data acquisition, annotation and analysis

For our pilot study, we implemented and tested the whole multimodal procedurę on one participant.

2.1. Participant

A 33 years old French typical right-handed male speaker (MO1) was recorded in the premises of the LORIA laboratory in Nancy. The speaker had previously filled out a document asking for his consent indicating the different steps of the recordings and the equipment used.

2.2. Experimental paradigm

MO1 was recorded during a narrative task, in an experimental situation, inspired by McNeill's protocol (1992). Several sequences from a cartoon of the series *Tweety and Sylvester* (1949, Warner Brothers) were presented to him. After viewing each sequence, MO1 had to narrate it immediately to an interlocutor. MO1 was filmed throughout the duration of the task. We cut the cartoon into 5 sequences, including the "strike" sequence frequently exploited by the gesture community and which was chosen for this study in order to present the processing chain used to study the synchronization of gestures and prosody. For this short paper, we will focus on the acoustic data and on the gestural data exported from the IMU (Inertial Measurement Units, see just below).

2.3. Recording procedures

MO1 was recorded with two different motion capture devices (mocap). The first device (see Figure 1, left) consists of an electromagnetic articulograph (EMA) to record the movements of both hands and speech articulators (lips, tongue, jaw), a microphone to capture the acoustic signal and a video camera placed facing the speaker to film the entire scene. The device is completed by a laryngograph, which records the activity of vocal fold vibrations and a breathing belt recording the subject's abdominal movements. The EMA is normally used to study the movements of the main articulators of speech, i.e. the lips, tongue and jaw. The different movements are recorded every 5 ms using sensors placed on these different articulators. An electromagnetic field inducing an alternating current in the sensors makes it possible to measure the distance between the sensors and the transmitters (absolute measurement). We diverted it from its original use by placing 3 sensors on each hand (6 sensors in total) and 6 remaining sensors on the face and tongue (the device being equipped with 12 sensors in total). The sampling frequency of the EMA is 300 Hz and the recording of the speech signal (16 bits, 16 kHz) is synchronized with the recording of the magnetic signals provided by the sensors. The second device (see Figure 1, in the middle) is composed of a suit that forms a serie of Inertial Measurements Units (IMU) and captures the body’s movements with 32 sensors located on the entire body (inertial units). The data is then visualized in 3D with the AXIS Neuron software (see Figure 1, right).

[1] Project funded by a grant awarded by the CNRS as part of the "Challenge Instrumentation aux Limites" call for projects in 2017 (Coordination: Christelle Dodane).

Figure 1: Speaker (MO1) recorded with two different mocap systems: the electromagnetic articulograph (EMA, on the left) and the IMU suit (in the middle) with the visualization of the latter in Axis Neuron (on the right).

2.4. Coding and processing of acoustic and prosodic data

The sound files were segmented with the Praat software (Boersma & Weenink, 2018) and result in a 6-line grid (called "tiers", see Figure 2). The “phoneme”, “syllable” and “word” tiers, were automatically segmented with the EasyAlign software (Goldman, 2011) and then manually corrected. The following three tiers were annotated manually. The "Prom" tier includes the annotation of perceptual salient syllables (prominences) following the procedure recommended by the *Rhapsodie* ANR prosodic coding protocol (Lacheret *et al.*, 2014). The strong prominences ("S") were coded by ear at a coding span of 5 seconds. Then the weak prominences ("W") were annotated. The prominences marked by a sharp rise in the fundamental frequency (F0) were annotated "S°°" or "W°°" and those marked by a smaller rise, "S°" and "W°". A fifth tier was added to manually annotate the different intervals corresponding to the points of inflection of the F0 and the value of the F0 corresponding to these points (upward contours, "M", downward contours, "D" and flat contours, " P "). And finally, a sixth tier was added containing the annotations of the different disfluencies, with the aim of comparing the disfluencies of stutterers and normo-fluent subjects, since one of the future applications of our project is the analysis of the coordination between gestures and speech in subjects who stutter.

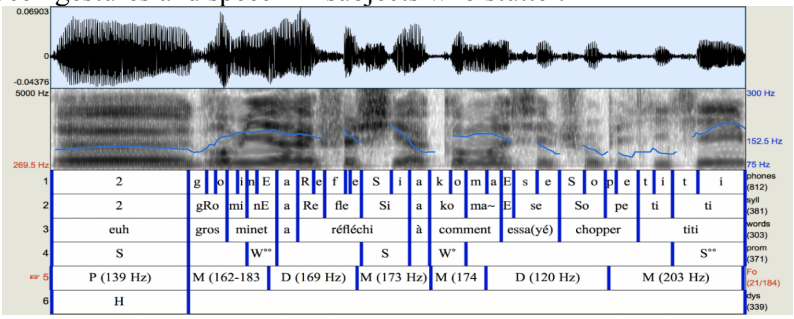

Figure 2: Segmentation grid of a statement produced by the speaker MO1, with six tiers (Phonemes, Syllables, Words, Prominences, F0, Disfluencies) provided by the Praat software.

The evolution of the F0 value (in Hertz) was then extracted automatically with the Praat software (every 10 ms for the F0 and every 10.666 ms for the intensity) and imported into the ELAN software (Sloetjes & Wittenburg, 2008) to be synchronized with gestural data. To obtain a stylization of pitch variations according to a tonal perception model, we used the "Prosogram" application from Mertens (2019).

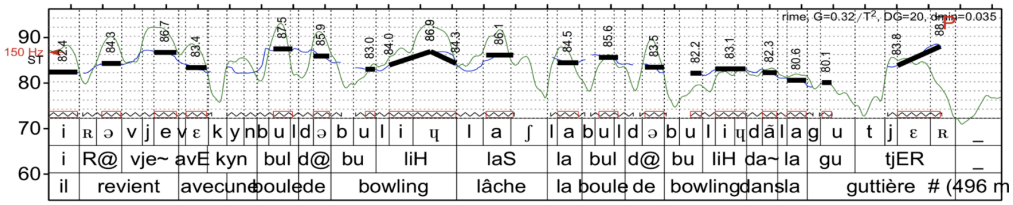

Figure 3: Prosogram of a statement produced by MO1, with in black, the automatic detection of the temporal evolution of pitch, and on the top, the targets in semitones.

2.5. Synchronisation of the recordings

The data retrieved from the three sources - video, audio, and motion capture (mocap) – had to be synchronized with each other since the recordings did not start at the same time. Synchronization was performed under ELAN in which we can integrate the audio, video and mocap sources (with a beep or manual clap at the beginning).

2.6. Sampling frequencies

In addition to synchronization, the frequency of each of these recordings is not the same, it is even different within the same audio source. Indeed, the sampling frequency of the pitch is 10 ms. In concrete terms, this means that the gap increases as time goes by. The sampling rate of the images in the video is 40 ms. The timespan in transcripts under ELAN or Praat is variable and can be done to the nearest millisecond. The timespan for the mocap (Inertial Movement Unit) is 16.5 ms. Four different frequencies coexist in the data, in increasing order: a millisecond for transcription under Praat and / or ELAN, 10 ms for the pitch, 16.5 ms for the mocap and finally the timespan of the video is 40 ms. Video serves us primarily as a visual synchronization element, the data are not processed on this visual basis. In any case, we quickly found gaps in the data. These gaps increase progressively, and vary according to the type of data. It was therefore necessary to re-sample continuously in order to calibrate and coordinate the data without creating false data.

2.7. Resampling method

As we wanted to avoid to create false data, by using interpolation for example, the principle of resampling consisted in aligning the data associated with a short timespan (10 ms) from the existing data associated with a longer span in frequency. Thus, the first four temporal values of the mocap (i / 0 ms, ii / 16.5 ms, iii / 33 ms, iv / 46.5 ms), were aligned with the pitch data associated with the first six values (i / 0 ms, iii / 20 ms, iv / 30 ms, vi / 50 ms). Step by step, every 16.5 ms, the mocap data were compared with the pitch data that corresponded to the closest temporal values. When the matching by resampling of the pitch and mocap data were done, we needed to compare this re-alignment of the data with the Praat transcripts. Each unit (word, syllable, phoneme) has a beginning and an end. These intervals do not correspond to a fixed timespan, they depend entirely on what has been produced by the speaker. The temporal values of word boundaries can be corrected based on the closest values in the mocap output (values are inferior to 8.25 ms, ie 16.5 / 2).

3 Results

Table 1 summarizes our main results for pitch. The speaker has approximately the same speech rate in both tasks. He has a larger speech range and a lower mean pitch with the IMU suit.

Mocap	Speech rate	Pitch range	Mean Pitch	Max pitch	Speech time	Phonation time	Pause time
EMA	9,09 syll/sec.	8,1 ½ tons	154 Hz	196 Hz	110,84 sec.	8,91 sec. (8,04 %)	101,93 sec. (91,96%)
Noitom Suit	9,2 (6,23) syll/sec.	9,2 ½ tons	129 Hz	169 Hz	71,85 sec.	55,52 sec. (77,28%)	16,33 sec. (22,72 %)

Table 1: measures based on the Prosogram application for the extract in which the motion capture was used along with the EMA and the IMU suit.

All the studies that have so far explored the relationship between prosody and gestures have followed the positions and movements of the hand according to an absolute frame of reference. Among the sets of devices used in this study, the EMA falls under this type of absolute reference framework. In order to record the co-verbal gestures, these sensors are placed on the hands only, providing data on the position and movement of the hands in a unique and absolute reference frame being located in the recording room. Note that the hands may be submitted to a movement from higher up (arms, shoulders) without having moved on their own, i.e. the consequence of a movement of the arm is measurable on the hand. With a device like the EMA, one cannot detect and analyze the movements of the other segments nor the movement of the hand itself. Thus, to find out what the movements of all the segments of the upper limb are like, we can use the data from the IMU. The IMU enables us to situate the gestures in as many intrinsic reference frames as there are segments: the position and the movement of each segment are given with respect to the adjacent and proximal segment. Thus, the movements of the arm are calculated according to the shoulder, those of the forearm, according to the arm, those of the hand

are determined relative to the forearm. It is therefore possible to measure which segment is moving and by which angle in the three dimensions of each one's own space. The results of the relations between prosody and gesture in these intrinsic frames of reference are presented below. We can thus follow the evolution of the pitch and its possible impact on one of the 8 degrees of freedom of the upper limb, distributed from shoulder to hand. To our knowledge, these links have never been made. A correlation (Bravais-Pearson) was established between the rising (N = 63; mean time = 90.40 ms) and descending (N = 111; mean time = 92.74 ms) ranges of F0 and the degrees of freedom of the three segments (arms, forearm and hand) of the right upper limb and shoulder for the IMU recording. The linear correlation coefficients range between 1 and -1. Notice that there is a strong affinity between two sets of variables when their value is between 0.8 and 1 or between -0.8 and -1. As we get closer to the value 0, the series are less, if at all, correlated. These results are presented in the two tables below.

F0 ⬈	Add/Abd Shoulder & F0⬈	Rot Ext/Int Arm & F0⬈	Exten/Flex Arm & F0⬈	Add/Abd Arm 1 F0⬈	Supi/Pro Forearm & F0⬈	Exten/Flex Forearm & F0⬈	Add/Abd Hand & F0⬈	Exten/Flex Hand & F0⬈
% corr. coef. 1> x >0,8 or -0,8> x >-1	55,56%	88,89%	82,54%	75,81%	74,60%	80,95%	68,25%	73,02%
% corr. coef. 0,8> x >0,6 or -0,6> x >-0,8	17,46%	3,17%	9,52%	8,06%	9,52%	6,35%	7,94%	9,52%
% corr. coef. 0,6> x >0,4 or -0,4> x >-0,6	6,35%	6,35%	1,59%	3,23%	6,35%	7,94%	9,52%	11,11%
% corr. coef. 0,4> x >0,2 or -0,2> x >-0,4	12,70%	1,59%	6,35%	8,06%	9,52%	3,97%	6,35%	3,17%
% corr. coef. 0,2> x >-0,2	7,94%	0,00%	0,00%	4,84%	0,00%	0,79%	7,94%	3,17%

Table 2: Percentages of the number of correlation coefficients per range of 0.2 between rising fundamental frequencies and each degree of freedom of the right upper limb. The set of gestural possibilities are defined by degrees of freedom from shoulders to hands included[2].

F0 ⬊	Add/Abd Shoulder & F0⬊	Rot Ext/Int Arm & F0⬊	Exten/Flex Arm & F0⬊	Add/Abd Arm 1 F0⬊	Supi/Pro Forearm & F0⬊	Exten/Flex Forearm & F0⬊	Add/Abd Hand & F0⬊	Exten/Flex Hand & F0⬊
% corr. coef. 1> x >0,8 or -0,8> x >-1	50,45%	78,18%	72,97%	73,87%	69,37%	75,68%	60,36%	70,27%
% corr. coef. 0,8> x >0,6 or -0,6> x >-0,8	18,02%	12,73%	13,51%	14,41%	16,22%	14,86%	19,82%	15,32%
% corr. coef. 0,6> x >0,4 or -0,4> x >-0,6	16,22%	5,45%	5,41%	6,31%	4,50%	3,60%	6,31%	6,31%
% corr. coef. 0,4> x >0,2 or -0,2> x >-0,4	11,71%	1,82%	5,41%	4,50%	6,31%	3,15%	6,31%	4,50%
% corr. coef. 0,2> x >-0,2	3,60%	1,82%	2,70%	0,90%	3,60%	2,70%	7,21%	3,60%

Table 3: Percentages of the number of correlation coefficients per 0.2 range between descending fundamental frequencies and each degree of freedom of the right upper limb[3].

Tables 2 and 3 indicate that the correlations between the variations of F0 and the degrees of freedom are very strong (at least 60% of the cases higher than a coefficient that is equal to or higher than || 0.8 ||). Moreover, these correlations distributed over all segments of the upper limb, are particularly important for the arm and decrease globally as we take the movement of the forearm and hand into consideration. Even if the high correlation rate remains present for these latter segments, we notice a decrease in the co-variation with the pitch for the distal segments, in particular with a shift towards lower values (between 0.6 and 0.4). In other words, the further one gets away from the bust, in terms of segments

[2] The set of degrees of freedom are defined in the position of the Vitruvian man (man standing with his palms facing forward, circumscribed in a circle, illustrated by Leonardo da Vinci). Abduction / adduction is a degree of freedom that moves a segment or a shoulder away or closer to the bust, in a frontal plane. The extension / flexion makes the segment pass behind or in front of the frontal plane, still in this general reference position of the Vitruvian body. The outer / inner rotation and the supination / pronation are degrees of freedom that turn the segment on itself, arm for the first, forearm for the second.

[3] The negative values of the correlation coefficients appear when for the same F0 slope, the pole of the correlated degree of freedom is of an opposite sign. Thus, when the correlation coefficient is negative, for a descending F0, then for example for the arm, its movement corresponds to a flexion (forward or upward). For a positive value, always with descending F0, the arm will have an extension movement (backward or downward).

(and not of distance), the less powerful this co-variation becomes. We don't know yet whether this anisotropy is structural or if it comes from a temporal shift due to the time needed for the movement of the arm to propagate towards the hand. In favor of this last hypothesis, the average duration of the variations of F0 is about 90 ms when the average duration of the gestures is about 150 ms. A gesture that begins from a proximal segment, cannot have fully developed over all the segments by the time the rise or fall of the fundamental frequency is reached. These questions explain a) the common structuration between prosody and gesture b) their synchronization c) the management of various temporalities.

4 Perspectives

The MODALISA project has reached its technological goal as we have now created a multimodal, multidevice platform in order to collect data on both speech and gesture as well as a methodology to process and analyze the multimodal data. The pilot study we presented in this paper indicates strong correlations between gestures and intonation when they are simultaneously performed by the speaker. The correlations are particularly strong for proximal segments. It would thus be particularly important to analyze head gestures (as advised by Bolinger, 1983). The advantages of the MODALISA platform are that we use MOCAP systems with different frames of reference (absolute (EMA)/intrinsic (IMU) and that we have the possibility to integrate articulatory gestures (EMA), respiratory movements (respiratory belt), vibratory movements of the larynx (laryngograph) with prosody and gesture. We aim to adapt the IMU suit to children's physiological constraints. We can also coordinate the various exported data with our annotations of the video-data on ELAN. The platform is used in various projects by our team to study how prosody and gesture synchronize across the life-span in typical and atypical populations. Our goal is to capture whether integration of polysemiotic resources is quantitatively or qualitatively different in children as their motoric, cognitive and linguistic skills develop and in adults as they reach old age.

Acknowledgments: The authors would like to thank Marjorie Bosqué and Cwiosna Roques for their annotation work on the data files.

References

Balog, H. and Brentari, D. (2008). The relationship between early gestures and intonation. *First Language*, 28, pp. 141-163.

Boersma, P., Weenink, D. (2009). Praat: doing phonetics by computer (Version 5.1.15) [Computer Program]. Consulté le 25.01.2017 de PRAAT: http://www.praat.org/

Bolinger, D. L. (1983). Intonation and gesture. *American Speech*, 58, 156-174.

Cruttenden, A. (1997). *Intonation* (second edition). Cambridge: Cambridge University Press.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*, London: John Murray.

Goldman, J.P. (2011). EasyAlign: an automatic phonetic alignment tool under PraatProceedings of InterSpeech, September 2011, Firenze, Italy.

Guellaï, Bahia, Langus, Alan et Nespor, Marina. (2014). Prosody in the hands of the speaker. Frontiers in Psychology. 5, 700, 1-8.

Iverson, J.M. and Goldin-Meadow, S. (1998). Why people gesture when they speak. Nature, 396, 228.

Kendon, A. (1980). Gesticulation and speech: Two aspexts of the process of utterance. In M. Key (ed.), "The Relationship of Verbal and Nonverbal Communication", The Hague: Mouton, 207-227.

Kendon, A. (1983). Gesture. Journal of Visual/verbal Languaging, 3 (1), 21-36.

Lacheret, A., Kahane, S., Beliao, J., Dister, A., Gerdes, K. et al. Rhapsodie : un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé. 4e Congrès Mondial de Linguistique Française, Jul 2014, Berlin, Allemagne. 8, pp.2675-2689, 2014.

Locke, J.L., Bekken, K.E., McMinn-Larsen, L. and Wein, D. (1995). Emergent control of manual and vocal-motor activity in relation to the development of speech. Brain Language, 51, 498-508.

McNeill, D. (1992). Hand and Mind. What Gestures Reveal about Thought. Chicago: The University of Chicago Press.

Mertens, P. (to appear early 2019). The Prosogram model for pitch stylization and its application in intonation transcription. In Barnes, J.A. & Shattuck-Hufnagel, S. (eds.). Prosodic Theory and Practice. Cambridge, MA: MIT Press.

Ouni, S., Mangeonjean, L., Steiner, I. (2012). VisArtico: a visualization tool for articulatory data. *Interspeech 2012*, September 9-13, 2012, Portland, OR, USA.

Roustan, B. & Dohen, M. (2010). Co-production of contrastive prosodic focus and manual gestures: temporal coordination and effects on the acoustic and articulatory correlates of focus. 5th International Conference on Speech Prosody (Speech Prosody 2010), May 2010, Chicago, United States, 100110:1-4, 2010.

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

Vaissière, J. (1997). Langues, prosodie et syntaxe. Revue Traitement Automatique des Langues, ATALA, 38.I : 53-82.